Merging Person-Specific Bio-Markers for Predicting Oral Cancer Recurrence Through an Ontology

Dario Salvi , Marco Picone, María Teresa Arredondo, María Fernanda Cabrera-Umpiérrez
Ángel Esteban, Sebastian Steger, and Tito Poli

***Abstract*—One of the major problems related to cancer treatment is its recurrence. Without knowing in advance how likely the cancer will relapse, clinical practice usually recommends adjuvant treatments that have strong side effects. A way to optimize treatments is to predict the recurrence probability by analyzing a set of bio-markers. The NeoMark European project has identified a set of preliminary bio-markers for the case of oral cancer by collecting a large series of data from genomic, imaging, and clinical evidence. This heterogeneous set of data needs a proper representation in order to be stored, computed, and communicated efficiently. Ontologies are often considered the proper mean to integrate biomedical data, for their high level of formality and for the need of interoperable, universally accepted models. This paper presents the NeoMark system and how an ontology has been designed to integrate all its heterogeneous data. The system has been validated in a pilot in which data will populate the ontology and will be made public for further research.**

***Index Terms*—Biomedical image processing, cancer, computer aided diagnosis, genetic expression.**

I. Introduction

CANCER is the second cause of death in western countries. Although current treatments can be effective, the main problem of cancer is its recurrence either locally or by distant metastases, which are difficult to predict and prevent. The oral squamous cell carcinoma (OSCC), focus of this research, accounts for the 5% of all cancers, and has a rate of 25–50% of recurrence in five years, 90% of which within two years from surgery [1]. In order to avoid relapses, adjuvant chemo- or radio-therapy treatments are usually administered to all patients during follow-up, even in absence of disease signs. These treatments are heavy and have strong side effects that may harm also patients who are in fact already completely recovered. Knowing in advance which patients have the higher risk of disease recurrence, would help to initiate adjuvant treatments only in a limited, high-risk subgroup. In addition, the early identification of a neoplastic recurrence during follow-up would allow starting an appropriate treatment in time.

The most classical method to predict OSCC recurrence is the tumor, node, and metastasis (TNM) staging, which is based mainly on the dimensional characteristics of the tumor and on the presence, number, and site of neck nodes metastasis. Unfortunately, its inadequacy is today recognized because of the uncertain behavior of squamous cancer, which can be sometimes very aggressive and others can metastasize slowly after surgery [2]. This uncertainty in progression has led researchers to seek a larger number of markers. Many clinical, histopathological, radiological, and genetic factors were studied, but none of the different groups taken, distinctly provides clinically applicable markers of tumor aggressiveness [3]. Given the multilevel nature of cancer (genes, cells, tissues, and organs), integration of the different groups of data are required. Whereas different reports are present in the literature on data integration and creation of standardized prognostic algorithms for bladder and breast cancer, nothing is available for head and neck cancer. To cover this lack, the NeoMark project was created.

NeoMark [4] is an European cofunded research project, which aimed at identifying the optimal set of patient-specific and disease-specific bio-markers with a high-predictive power for the case of OSCC cancer.

The NeoMark strategy is designed to be integrated into normal staging and follow-up protocols. Patients are assessed before treatment and, at the time of remission, a wide range of data is collected including clinical observations, radiologic, and genomic data. A set of relevant markers expressed only in presence of the disease is then selected and relapse probability is estimated. If the same set of bio-markers appears during postremission follow-up, it would show a high probability of relapse, advising early intervention.

This strategy is supported by an Information and Communications Technology (ICT)-based system which allows physicians to

1) administer patients;
2) upload clinical data including histological information, surgery evidence, and risk factors;

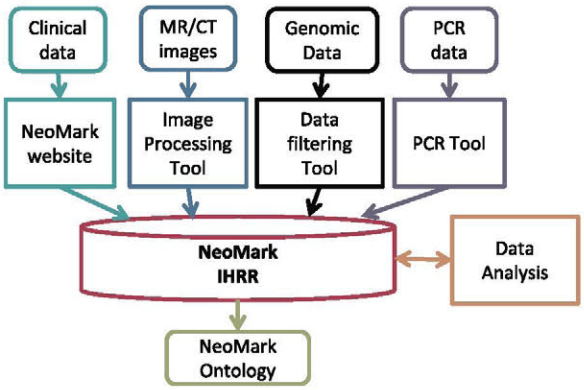

Fig. 1. Architecture of the NeoMark system.

3) analyze jointly multiple images from MR/CT scans;
4) analyze gene expressions by means of microarrays and a mobile PCR system;
5) receive indicators of the probability of relapse for supporting the clinical decision during the follow up;
6) download anonymized data for further research and statistics.

The following sections describe the details of the NeoMark system, how an ontology for integrating all the data collected in the system has been designed, some results of the pilot we ran to prove its effectiveness and the conclusions and some suggestions for future developments.

II. Methods

A. NeoMark System

The NeoMark system was designed as a flexible, user-friendly, service-oriented system. The system comprises the following modules, as depicted in Fig. 1.

1) Integrated Health Record Repository (IHRR): It is a central database that collects all the patients data in a central entity anonymously (only a unique identifier is kept for each patient). A web-based data entry provides all the functionalities to add, manage, and review patients data. In order to link patients clinical data with their sensitive information, a stand-alone application has been developed to manage this link with the database located in the hospital network.

Additionally to the web data entry, the IHRR provides a remote interface to other three dedicated applications that upload patients data, and to a data analysis module that is in charge of creating patient-specific models for estimating cancer recurrence. These external tools are detailed in the following sections.

2) Image processing tool (IPT): It is used to extract meaningful numeric features of tumors and suspicious lymph nodes from CT, CT with contrast, MR T1 TSE, and MR T2 TSE head and neck images. Prerequisite for the feature extraction is an image fusion of the CT and MRI scans provided by the means of image registration as well as the segmentation of the region of interests (tumors and suspicious lymph nodes).

For image registration, the IPT deploys a fully automated mutual information-based rigid registration method [5]. It is robust against truncation and imaging artifacts as they are typically present in clinically acquired medical images. In order to satisfy the rigidity assumption, not only in the head but also in the deformable neck, the patients had been placed in the same position during image acquisition using the different imaging modalities. In a second step, the operator selects relevant lymph nodes by clicking into the approximate center. Starting from there, a radial ray-based segmentation technique is used to segment the lymph nodes automatically [6]. Due to the unpredictability of the shape, appearance, and surrounding tissue of the tumor, an interactive approach is used for tumor segmentation [7].

In the third phase, from the segmentation process some geometric and texture-based features are automatically extracted, including volume, 3-D axes, contrast take-up rate and water content. Finally, the location and the amount of infiltration of the surrounding tissue are estimated manually. All these features, including images, are then uploaded to the IHRR.

3) Genomic data filtering tool: Blood samples and tumor samples are analyzed with a standard microarray scanner which produces a feature extraction (FE) file with the extracted values. A FE file is a tab delimited text file, which contains Log2-ratio values as well as raw intensity data, background information, metadata on the experiment and on the scanning settings and annotations to identify genes. In order to reduce the amount of collected data, we only consider as relevant information the feature name, probe name, gene name, systematic name, description, and Log2-ratio. The Genomic data filtering tool discards irrelevant data and low quality or duplicate features producing as output a cleaner and lighter file containing only relevant fields for the analysis. These new files are eventually assigned to the specific patient and sample and uploaded to the common repository, while the original copy of the FE file is kept in the local hospital repository.

4) qRT-PCR tool: Within the project, a portable, real time, low-cost quantitative real time polymerase chain reaction (qRT-PCR) tool has been developed. Its objective is to serve as a lab-on-a-chip alternative to microarrays for examining patient's RNA extracted from lymphocytes. The tool analyzes a set of predefined genes (up to 20) and reports their expression value in relation with a housekeeping gene. It is composed of a core silicon chip containing heaters and thermal sensors, a cooling fan, and an optical system that detects the fluorescence intensity of the monitored reaction. The practitioner, after preparing the chips with genetic material, operates a SW tool that controls the device, clears and filters data, and sends it to the IHRR.

5) Data analysis: It is a module in charge of creating the patient-specific estimation model of cancer recurrence. Its role is twofold. As soon as the patient reaches remission, an initial profile of the patient is created by analyzing clinical, imaging, and gene expression data. The objective of this first step is to stratify patients into two classes: remittents and nonremittents. A feature selection is executed for the purpose: outliers are detected, missing values are handled, and redundant features are discarded. The adopted feature selection mechanisms are significance analysis of microarrays (SAM) [8] and correlation-based feature subset selection (CFS) [9]. The reduced set of features is then fed to a set of classification algorithms which have been trained carefully.

The second role of the module is to model the evolution of the disease during the whole follow-up period in order to early identify potential relapses. The following classifiers are employed: bayesian networks (BN), artificial neural networks (ANN), support vector machines (SVM), decision trees (DT), and random forests (RF). In order to avoid overfitting, a tenfold cross validation is also performed.

B. NeoMark Ontology

The IHRR uses a standard relational database for storing and retrieving data. Nevertheless, one of the objective of the project was to provide the data in a semantic structure by means of an OWL ontology. For this reason, a component of the IHRR is dedicated to generating RDF instances of the data according to a specific model: the NeoMark ontology [10].

The objectives of the ontology are as follows:

1) to represent the domain of the project and documenting it;
2) to link the NeoMark data to other domains and ensure interoperability by including established existing ontologies;
3) to share the data created within the pilot with a formal semantic.

The NeoMark ontology was built starting from the existing database included in the IHRR. We developed a tool, based on the work shown in [11], that translates tables, columns, and foreign keys of a standard relational database into classes, data properties, and object properties. We run the tool on the NeoMark database and we obtained a first approximation of the ontology, which comprised all the concepts and all the relationships that were in use within the system, but represented in a poorly structured way and with almost no explicit semantic. Since the NeoMark domain integrates several types of heterogeneous data, we decided to base it on an upper ontology. The adoption of the principles of the open biological and biomedical ontologies (OBO) foundry [12] was the natural choice given the domain of our application. We imported the basic formal ontology (BFO) and the relation ontology (RO) and reordered the concepts of the automatically generated ontology on top of them. The result was a much more formal ontology, but was still lacking interoperability with other existing initiatives. For importing existing ontologies, we adopted the recommended approach of the OBO foundry: the minimum information to reference an external ontology term (MIREOT) [13]. More concretely, we used the OntoFox tool [14], which implements and expands the MIREOT approach.

We analyzed the ontologies currently deployed in the OBO foundry to find reusable concepts. Eventually, the following terms were imported:

1) from the ontology of biomedical investigations (OBI), the terms "patient role," "lymph node," "gene list," and "disease";
2) from the ontology for general medical science (OGMS), the "clinical finding," "prognosis," and "treatment" taxonomies;
3) from the human disease ontology (HDO), the malignant neoplasm of lip, oral cavity, and pharynx taxonomy.

The integration of these ontologies was simple because both OBI and OGMS are based on the BFO ontology, while HDO was included by making equivalent the term "disease" between HDO and OBI.

The existing ontologies were imported including all intermediate terms between classes and computing all axioms recursively. A further process was setup to prune irrelevant terms from the current model and reorganize the terms that were not included in the imported ontologies, but that were still necessary to represent the NeoMark domain. Following, we show the root class of the added taxonomies and their direct superclasses in the merged ontologies:

1) ``tumor finding'' taxonomy (subclass of ``clinical finding'');
2) ``lymph node finding'' taxonomy (subclass of ``clinical finding'');
3) ``patient quality of life evaluation'' taxonomy (subclass of ``physical examination finding'');
4) ``patient risk factor'' taxonomy (subclass of ``clinical finding'');
5) ``microarray data'' (subclass of ``laboratory finding'');
6) ``qPCR data'' (subclass of ``laboratory finding'');
7) ``gene list''(subclass of ``data set'');
8) ``gene ontology term'' (subclass of ``data set'');
9) ``recurrence prediction'' taxonomy (subclass of ``prognosis'');
10) ``surgical treatment'' taxonomy (subclass of ``treatment'');
11) ``nonsurgical treatment'' taxonomy (subclass of ``treatment'').

In addition to specific classes, some *ad hoc* relationships had to be added to the model. All relationships were derived from the relation ontology. The relations, described using Manchester OWL syntax, are shown as follows:

1) ``patient role'' contains only ``lymph node,''
2) ``patient role'' contains exactly 1 ``malignant neoplasm of lip, oral cavity and pharynx,''
3) ``malignant neoplasm of lip, oral cavity and pharinx'' contained_in exactly 1 ``patient role,''
4) ``lymph node'' contained_in exactly 1 ``patient role,''
5) ``lymph node finding''is_about exactly 1 ``lymph node,''
6) ``tumor finding'' is_about exactly 1 ``malignant neoplasm of lip, oral cavity, and pharynx,''
7) ``clinical finding'' is_about exactly 1 ``patient role,''

8) ''prognosis'' is_about exactly 1 ''patient role,''
9) ''patient role'' participates_in min 1 ''treatment,''
10) ''treatment'' has_participant exactly 1 ''patient role,''
11) ''treatment'' precedes only ''post treatment,''
12) ''post treatment'' preceded_by some ''treatment,''
13) ''surgical procedure'' precedes some ''surgical reconstruction,''
14) ''surgical reconstruction'' preceded_by some ''surgical procedure,''
15) ''gene ontology term list'' has_part only ''gene ontology term,"
16) ''gene ontology term'' part_of only ''gene ontology term list,''
17) ''microarray input gene list'' is_about exactly 1 ''patient role,''
18) ''microarray output gene list'' is_about exactly 1 ''patient role,''
19) ''gene ontology term'' contains exactly 1 ''gene ontology term gene list,''
20) ''gene ontology term gene list'' contained_in exactly 1 ''gene ontology term''.

III. Results

The NeoMark system was validated in a pilot which involved 86 OSCC patients treated with surgery with more than 12 months of follow-ups in three clinical centers in Italy and Spain.

From the technical perspective, the maturity, the usability, and the perceived usefulness were validated. The maturity of the system was evaluated by setting up an issue tracker (GForge) that was used by the clinicians during the study. In total, 58 technical problems were reported and 33 new features were requested. At the end of the study, no reported errors were still open and only a minority of requests for changes, on which no common agreement was reached, were still open.

For validating usability and usefulness, a specific questionnaire was sent to 24 respondents who were using the system in the clinical centers (radiologists, biologists, maxillofacial surgeons, and laboratory technicians). The questionnaire was designed with cross-check questions expressed in positive and negative answers. The aggregated results show positive feedback with regards to easiness of use and user friendliness (55% of respondents were positive) and usefulness (67.2% of respondents were positive). The validation of the developed tools, especially the image processing tool and the qRT-PCR tool demonstrated the possibility to improve the decision-making process of physicians by enhancing and speeding up some critical diagnostic exams and by integrating data through an accurate and usable software platform.

From the clinical perspective, risk prediction was validated, both as risk stratification of patients at diagnosis (baseline) and recurrence risk evaluation during follow-up. Clinicians validated the risk prediction results via vis-á-vis the real evolution of the disease in subsets of patients, and verified in parallel the factors and genomic markers identified by NeoMark against the most recent findings from the scientific research.

The classification performances per type of input data can be summarized as follows [15].

1) For the clinical data, accuracy is around 78%.
2) For the imaging-only data, accuracy is around 88%.
3) For the tissue genomic data, accuracy is around 88%.
4) The best performing algorithm for the cases aforementioned is the CFS combined with SVM classifier.
5) For the blood genomic, the most adequate classification scheme was ANN without any previous feature selection with an accuracy of approximately 96%.

When combining these classification schemes into a consensus classifier, accuracy reaches approximately 92% with 83% sensitivity and 100% of specificity.

The NeoMark data analysis also allowed us to select a better set of genes with higher prediction power than already identified in the literature, particularly.

1) From an initial set of 45015 genes, our best selection includes a mix of 37 genes. The best results are achieved when the genes identified from data analysis are combined with the ones in the literature. An optimized selection includes 9 genes from the literature and 11 found within NeoMark.
2) A minimized set of genes with an acceptable accuracy includes 9 genes, 5 of which were identified by the NeoMark system.

Regarding the evaluation of the NeoMark ontology, it was validated in different steps throughout its development. The evaluation of an ontology is typically carried out from two points of view: the technical evaluation is performed by developers and the users evaluation is done by experts of the domain [16]. In NeoMark, the technical evaluation for consistency was done with standard reasoners included in Protégé. Users' evaluation was performed in this way: while the terms were added or pruned from the existing taxonomies, five experts of the project, doctors, and biologists, helped us to assess the completeness and coherence of the model. These evaluations allowed us to correct inconsistencies, predominantly during the conceptualization phase and avoided errors to propagate. The current version of the ontology is publicly available at the National Center for Biomedical Ontology BioPortal[1].

We are currently completing the integration of the IHRR data with the NeoMark ontology and performing the last adaptations to guarantee that all the data of the pilot are correctly represented in the model. As soon as the the ontology export feature of the IHRR is finalized, all the data of the pilot will be made available as an OWL ontology to be downloaded at the NeoMark website[2]. We are also planning to integrate a SPARQL engine into the web interface of our system to make it easier for researchers to query for relevant data.

[1] http://purl.bioontology.org/ontology/NeoMark
[2] http://www.neomark.eu

IV. Conclusion and Future Work

This paper has shown a novel ICT-enabled cancer recurrence prediction method, a system that implements the method for the OSCC case and an ontology that models all the data managed by the system. The clinical validation has shown that NeoMark can enable the identification of integrated and reduced sets of markers. Almost all the clinical and imaging identified markers are coherent with the most recent findings of research in this field, while additional factors, especially related to tissue and blood genomic, were identified by the NeoMark. If NeoMark-specific genes are combined with most performing genes extracted from the literature, the accuracy of the prediction tool increases considerably.

The validation of the system, although generally positive, showed that some improvements are still necessary.

Regarding technical aspects, some fixes on the user interfaces are needed and the response time of the system must be reduced considerably. From the clinical point of view, risk prediction should be further assessed through a longer and wider clinical study, in order to assess how valuable it can be for decision making. At the present stage, the indications of NeoMark risk prediction can be helpful in triggering additional attention to aspects which may be not visible, especially in radiology and in follow-up visits. Currently, the system works as a sentinel, but it does not provide insights about the possible mechanisms through which the recurrence occurs. Its use, though, can be extended as an instrument for further research, e.g., combined with retrospective studies on tissue samples available in tumor banks.

Regarding the ontology, as it was based on the OBO standard upper ontologies, its future revisions could be easily extended with new OBO ontologies as substitutes of manual added terms and relations.

Acknowledgment

The authors would like to thank the whole NeoMark Consortium for their valuable contribution for the realization of this work.

References

[1] P. Boyle and B. Levin, "World cancer report 2008," International Agency for Research on Cancer, Lyon, France, Rep., 2008. Available: http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/wcr_2008.pdf

[2] J. A. Woolgar, "Histopathological prognosticators in oral and oropharyngeal squamous cell carcinoma," *Oral Oncol.*, vol. 42, no. 3, pp. 229–239, Mar. 2006.

[3] R. d. e. C. Lindenblatt, G. L. Martinez, L. E. Silva, P. S. Faria, D. R. Camisasca, and S. d. e. Q. Lourenco, "Oral squamous cell carcinoma grading systems—Analysis of the best survival predictor," *J. Oral Pathol. Med.*, vol. 41, no. 1, pp. 34–39, Jan. 2012.

[4] M. Picone, S. Steger, K. Exarchos, M. Fazio, Y. Goletsis, D. I. Fotiadis, E. Martinelli, and D. Ardig, "Enabling heterogeneous data integration and biomedical event prediction through ict: The test case of cancer reoccurrence," in *Software Tools and Algorithms for Biological Systems* (Advances in Experimental Medicine and Biology Series 696), H. R. R. Arabnia and Q.-N. Tran, Eds. New York: Springer-Verlag, 2011, pp. 367–375.

[5] S. Steger and M. Keil, "Automated initialization and region of interest detection for successful head registration of truncated ct/mr head & neck images," in *Proc.10th IEEE Int. Conf. Inf. Technol. Appl. Biomed.*, 2010, pp. 1–5.

[6] S. Steger and G. Sakas, "Image gradient based shape prior for the segmentation of not that spherical structures," in *Proc.9th IEEE Int. Symp. Biomed. Imag.*, May 2012, pp. 1252–1255.

[7] S. Steger and G. Sakas, "FIST: Fast interactive segmentation of tumors," in *Abdominal Imaging Computational and Clinical Applications* (Lecture Notes in Computer Science Series 7029), H. Yoshida, G. Sakas, and M. Linguraru, Eds. Berlin/Heidelberg, Germany: Springer-Verlag, 2012, pp. 125–132.

[8] V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc. Natl. Acad. Sci. USA*, vol. 98, pp. 5116–5121, Apr. 2001.

[9] M. A. Hall, "Correlation-based feature selection for discrete and numeric class machine learning," in *Proc. 17th Int. Conf. Mach. Learn.* San Francisco, CA: Morgan Kaufmann, 2000, pp. 359–366

[10] I. Martín, M. Ortega, D. Salvi, Á. Esteban, and M. Picone. (2011). "Neomark internal report 4.1, ontologies and semantic integration," Tech. Rep. [Online]. Available: http://lnx.neomark.eu/portal/images/articles/pdf/public_deliverables/ir 4.1_ontologies_semantic_integration.pdf

[11] Z. Xu, S. Zhang, and Y. Dong, "Mapping between relational database schema and owl ontology for deep annotation," in *Proc. IEEE Web Intell.ACM Int. Conf.*, Dec. 2006, pp. 548–552.

[12] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, the OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The obo foundry: Coordinated evolution of ontologies to support biomedical data integration," *Nature Biotechnol.*, vol. 25, no. 11, Nov. 2007.

[13] M. Courtot, F. Gibson, A. L. Lister, J. Malone, D. Schober, R. R. Brinkman, and A. Ruttenberg, "Mireot: The minimum information to reference an external ontology term," *J. Appl. Ontol.*, vol. 6, no. 1, pp. 23–33, 2011. DOI: 10.3233/AO-2011-0087. [Online]. Available: http://iospress.metapress.com/content/h54m2237310v13x1/

[14] Z. Xiang, M. Courtot, R. R. Brinkman, A. Ruttenberg, and Y. He, "Ontofox: Web-based support for ontology reuse," *BMC Res. Notes*, vol. 3, 2010. Available: http://www.biomedcentral.com/1756-0500/3/175

[15] K. Exarchos, Y. Goletsis, T. Poli, and D. Fotiadis, "Gene expression profiling towards the prediction of oral cancer reoccurrence," in *Proc. 33rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, Aug. 2011, pp. 8307–8310. Available: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?reload = true&arnumber=6092048&contentType=Conference+Publications

[16] A. Gómez-Pérez, M. Fernández-López, and Óscar Corcho, *Ontological Engineering with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web*. London, U.K.: Springer-Verlag, 2004, 420 pp. Available: http://www.springer.com/computer/information+systems+and+applications/book/978-1-85233-551-9